A Resume of Ph.D. Thesis entitled

ANALYSIS OF UNSTRUCTURED DATA USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Submitted to:

Swami Ramanand Teerth Marathawada University, Nanded

for the Degree of
Doctor of Philosophy (Ph. D) in
Computer Science
Under the Faculty of Science and Technology

Researcher

Miss Pooja Sanjay Malhare

Research Supervisor

Dr. Purushottam Anandrao Kadam
Assistant Professor
Department of Computer Science
Institute of Technology and Management VIP Road Nanded

Research Centre: School of Computational Sciences, SRTM University, Nanded

December 2024

Index

Sr No	Contents	Page No
01	Introduction	01
02	Review of Literature	05
03	Research Gap	06
04	Aims and Objectives of Research	07
05	Importance of Research	08
06	Scope of Research	08
07	Hypothesis	09
08	Research Methodology	10
09	Proposed Research Plan	12
10	Expected Outcomes	13
11	Limitations	13
12	References:	14

Title of the Research:

ANALYSIS OF UNSTRUCTURED DATA USING ARTIFICIAL INTELLIGENCE TECHNIQUES

1 Introduction:

In today's digital age, organizations are generating vast amounts of unstructured data, including text, images, audio, and video. This data is scattered across various sources, such as social media, emails, documents, and sensors. Traditional data analysis techniques are insufficient to extract insights from this unstructured data, which is why artificial intelligence (AI) techniques have become essential.

AI-powered tools can analyze vast amounts of unstructured data, identify patterns, and provide actionable insights. Techniques such as natural language processing (NLP), machine learning (ML), and deep learning (DL) can be applied to unstructured data to uncover hidden relationships, sentiment, and meaning.

A vast majority of the data now a day in digital world is of unstructured type. Unstructured data, also known as unstructured information, refers to information that lacks a predefined data model or is not organized in a specific way. Unstructured data, also known as unstructured information, refers to information that lacks a predefined data model or is not organized in a specific way. Unstructured data is the data, which cannot be stored in a relational databases (RDBMS) traditional structure. It is sometimes referred to as qualitative (estimated) data. We can develop meaning from it, but it may also be very unclear and difficult to analyze. Unstructured data does not follow to a specific data model, it does not have simple recognizable structure such that it cannot be used by computer software effortlessly. It does not have pre-defined model or structure, therefore it does not proper in relational database.

The file type is known when working with unstructured data, but the file content is unknown. Such files' contents are self-explanatory. They are difficult to analyze because they do not fit in a relational database. According to IDC (International Data Corporation), 80% of the world's data will be unstructured by 2025. Organizations have recently begun to pay attention to unstructured data, in particular the wealth of information that is stored in unstructured data and that is concealed behind unstructured data, and they now want to take advantage of it.

In order to derive insightful information from unstructured data and make pertinent business decisions, organizations must do this analysis. The decisions made based on these results will influence how customers feel about future planning, help to understand customer needs, and help to identify contributions that will better meet those needs. Machine learning algorithms and natural language processing techniques are used to analyze unstructured data to train a model to work for a specific need.

2. Sources of unstructured data

Broadly there are two main sources of unstructured data, the machine generated and human generated.

1. Machine generated unstructured data:

a. Satellite images (google maps to detect live traffic)

b. Data from medical imaging devices (ECG, MRI, etc.,)

c. CCTV surveillance data

d. Atmospheric data (temperature, humidity level, speed of wind etc.,)

e. Radar data (live tracking of flights)

2. Human generated unstructured data

a. E-mails (text, images, gif etc.,)

b. Text messages (text, images, gif, multimedia etc.,)

c. Social media data (tweets and post with text and multimedia)

d. Audio recording

e. Office documents (word, excel, pdf etc.,)

f. Images (photos taken from smartphone and digital camera)

3. Analysis of unstructured data

We must first determine which source data is necessary for the data analytics before we can move forward.

Unstructured data can be found in a variety of formats, including webpages, tweets, posts, videos, audio, and images; we must only analyze the data that is pertinent to our particular need. While there are various sources of data accessible, it is vital to choose the ones that are most relevant and applicable to our particular requirements.

It may be necessary to remove unnecessary data or anything that is not fit for purpose. It is also necessary to choose appropriate technology and tools for gathering, refining, storing, and managing data. Unstructured data is saved in its original format, along with the corresponding metadata.

Once all of these elements are established, we can devise a methodology for processing and analyzing the unstructured data. Now, let us examine the potential steps involved in analyzing unstructured data.

1. Metadata

Data about data is known as metadata, means the data itself provide information related to its data. It is very vital to store, manage and analyze the unstructured data. Object storage systems enable the storage of vast quantities of disorganized data, which is typically written just once and then accessed multiple times. What is Object storage exactly? In brief, it is a storage for unstructured data that removes the scaling limitation of traditional file storage. Unlimited scale is the motive that object storage is the storage of cloud.

The purpose of object storage is to store things like Facebook photos and videos, Gaana music files (songs), or files in online collaboration services like outlook and google drive.

The storage mechanism involves the creation of objects that consist of data, metadata, and a unique identifier. There are various metadata fields that can be filled in for most file types. Each image we take with a digital camera or smartphone has metadata like the date, time, filename, and location attached to it. Metadata such as the blog title, writer of the blog, URL of blog, blog issuing date, labels (tags), and category are included in every blog post. Metadata on a page incorporates things like the page title, URL of the page, page portrayal, and symbol.

We can create additional custom metadata fields based on your needs to describe the nature or contents of the unstructured data in addition to these standard fields. In this way, metadata can make further search and analysis easier.

Each enterprise must establish its own metadata standards because there are presently no industry-wide ones. Effective metadata use facilitates data organization, automation, policy enforcement, and data visibility. Even though it's ideal to attach metadata at the time the data is created, sometimes it needs to be done after the fact.

2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of computer science, and more specifically, artificial intelligence (AI), that focuses on enabling computers to comprehend written and spoken language in a manner similar to humans.

Natural Language Processing (NLP) is a machine learning technique that helps to analyze the meaning of unstructured textual data. NLP attempts to replicate the human mind's ability to process natural languages like Hindi, Gujarati, English, and others.

Even when documents do not adhere to a standard template, NLP can determine the meaning of text data in a context.

Let's examine some NLP models for handling unstructured text.

- Bag of Words: This technique simply involves counting the frequency of a word or phrase in the text, without considering its semantic context.
- Tokenization: It is a process used to divide strings into words and sentences, while disregarding punctuation, commonly used in natural language processing.
- Stop words removal: It is a technique that involves removing articles and prepositions from the text that do not add any significant value to the NLP process. These words include "a," "the," "and," and "to." A basic list of stop words is usually used to start the analysis, which can be refined depending on the specific goal of the analysis

- Stemming: It is a natural language processing technique that involves adding a prefix or suffix to a word in order to group different variations of the same word together. For instance, "sitar" and "sitarist" can be organized using stemming.
- Lemmatization: It is a process that involves converting words to their dictionary form, also known as their "lemma." This involves removing tenses, so words like "teaching" and "taught" both become "teach." Synonyms are also combined during lemmatization; for example, "home" becomes "house." Lemmatization considers the context of the word, as the same word may have multiple meanings depending on the context in which it is used. Topic modeling: A text-mining tool called topic modelling can help to identify themes in the text and locate word clusters associated with several themes. The words that appear frequently in a document can be used to identify its themes and meaning. In order to determine the topic of unstructured data, topic modelling involves adding up words and combining related patterns. A topic model group can determine whether a pattern is similar by looking at factors like word frequency and word distance, as well as the words and expressions that appear the most frequently.

3.Image Analysis

Let's focus on images now since unstructured data not only includes text but also a large portion of image content. We need to realize that images contain a wealth of information. If we have an image of a student learning a subject in a classroom, for instance, we can use image analysis to determine the student's behavior throughout the lesson in order to identify human behaviour by analyzing their gestures.

New systems have been created that can retrieve images by analyzing unstructured data, such as MRI images that correspond to a specific brain volume or X-rays of the backbone that match a particular backbone image. These systems utilize feature extraction and similarity matching techniques to identify similar images based on an input image. It is feasible to have automatic or intelligent vehicles on the road thanks to AI-based image analysis because they can identify objects on the road and determine their locations.

4. Data Visualization

Data visualization is a technique for putting data in a graphical format so that it is much easier to understand. Data that has been graphically visualized is easier to understand, encourages interaction, and reveals a wealth of new information. Data visualization makes it possible for more people to understand complex data structures that would otherwise be impossible to understand.

For example, I am having a Samsung Smartphone, a feature called "Digital Wellbeing" displays an app dashboard where users can see how long each application was open. By swiping to different screens, users can also view breakdowns by day, by hour, and by app, all of these is displayed in a well design data visualization manner. I can easily understand how much time spent with a certain app through the week. The most appropriate visualization techniques for the application and the intended audience can be

used to present unstructured data that has been processed using methods like topic modelling, sentiment analysis, information extraction from images and data visualization.

4. Challenges with unstructured data since the diagram makes it clear that 80% of data is unstructured, it is essential to understand what that data means, it stores wealth of information in unstructured data and it is hidden. This can only be done by analyzing the data, which is aided by artificial intelligence and machine learning. But, unstructured data presents a variety of challenges, from collection to storage to decision-making. Such data cannot be analyzed easily with the present database as most of the current data analytics databases are intended for structured data. Consequently, data analytics engineers have to look for the newer methods to locate, extract, organize and store
unstructured data.

Here is a list of the difficulties encountered when working with unstructured data.

- Conventional system such as RDBMS is not capable to analyze this data.
- Since unstructured data grows continuously, it has a very high volume of data, therefore the larger the dataset, the harder it is to store and manage, to deal with such data, we need systems that can process large volume of data efficiently.
- When we analyze the unstructured data, the relevance is one of the biggest issues since the data analytics models cannot make distinction between action and relationship.
- Unstructured data is frequently not available in high quality, particularly when it comes to images and videos.

5. Conclusion

We should make use of contemporary techniques like artificial intelligence and machine learning to analyses unstructured data in order to extract the wealth of information hidden within it because the nature of unstructured data makes it difficult to analyses and conventional database analytics would not be able to do it. Data analytics engineers are currently creating different algorithms and testing them on various datasets to gain insightful information. My own research work is related to this, where I proposed to develop a model that identifies gestures to determine human behaviour from unstructured
data.

2) Review of Literature:

The rapid growth of unstructured data from various sources such as social media, emails, documents, and sensors has created a need for effective analysis techniques. Artificial intelligence (AI) techniques have emerged as a promising solution for analyzing unstructured data.

Natural Language Processing (NLP)

NLP is a subfield of AI that deals with the interaction between computers and humans in natural language. NLP techniques such as tokenization, stemming, and lemmatization are used for text

preprocessing (Kumar et al., 2019). Sentiment analysis, entity recognition, and topic modeling are popular NLP applications for unstructured data analysis (Liu et al., 2019).

Machine Learning (ML) and Deep Learning (DL)

ML and DL are widely used for unstructured data analysis. Supervised learning techniques such as support vector machines (SVM) and random forests are used for classification and regression tasks (Kotsiantis et al., 2007). DL techniques such as convolutional neural networks (CNN) and recurrent neural networks (RNN) are used for image and text analysis (LeCun et al., 2015).

3) Research Gap:

Methodological Gaps

1. *Lack of standardized frameworks*: There is a need for standardized frameworks for analyzing unstructured data using AI techniques.
2. *Limited generalizability*: Many AI models are developed for specific domains or datasets, limiting their generalizability to other contexts.
3. *Insufficient attention to data quality*: Data quality issues can significantly impact the accuracy of AI models, but this is often overlooked in research.

Technical Gaps

1. *Handling high-dimensional data*: Unstructured data can be high-dimensional, making it challenging to analyze using traditional AI techniques.
2. *Dealing with noisy or missing data*: Unstructured data can be noisy or incomplete, requiring robust AI techniques to handle these issues.
3. *Scalability and efficiency*: AI models can be computationally expensive, requiring efficient and scalable solutions for large datasets.

Domain-Specific Gaps

1. *Limited research in certain domains*: Certain domains, such as healthcare or finance, require specialized AI techniques and expertise, but research in these areas is limited.
2. *Need for domain-specific AI models*: AI models developed for one domain may not be transferable to another domain, highlighting the need for domain-specific models.
3. *Insufficient attention to ethical considerations*: AI research often overlooks ethical considerations, such as bias and fairness, which are critical in certain domains.

Interdisciplinary Gaps

1. *Need for collaboration between disciplines*: Analyzing unstructured data using AI techniques requires collaboration between computer science, statistics, and domain-specific experts.
2. *Limited understanding of human-AI interaction*: As AI becomes more pervasive, there is a need to understand how humans interact with AI systems and how to design more effective human-AI interfaces.

3. *Insufficient attention to societal implications*: AI research often overlooks societal implications, such as job displacement and bias, which require interdisciplinary collaboration to address.

4) Aims and Objectives of Research:

Primary Aims

1. *Extract Insights*: Unlock hidden patterns, relationships, and meaning from unstructured data using AI-powered tools.
2. *Improve Decision-Making*: Provide actionable recommendations to stakeholders by analyzing unstructured data and identifying trends, opportunities, and risks.
3. *Enhance Customer Experience*: Analyze customer feedback, sentiment, and behavior from unstructured data sources to improve customer satisfaction and loyalty.

Secondary Aims

1. *Identify Areas for Process Improvement*: Analyze unstructured data to identify inefficiencies, bottlenecks, and areas for process optimization.
2. *Develop Predictive Models*: Use machine learning and deep learning techniques to develop predictive models that forecast future trends and outcomes.
3. *Detect Anomalies and Risks*: Identify potential risks, anomalies, and threats from unstructured data sources to improve risk management and compliance.

Specific Aims

1. *Sentiment Analysis*: Analyze customer sentiment from social media, reviews, and feedback forms to understand public opinion and perception.
2. *Entity Recognition*: Identify and extract specific entities such as names, locations, and organizations from unstructured data sources.
3. *Topic Modeling*: Identify underlying themes and topics from large volumes of unstructured data using techniques such as Latent Dirichlet Allocation (LDA).
4. *Intent Analysis*: Analyze customer intent from unstructured data sources such as chat logs, emails, and social media to improve customer service and support.

Strategic Aims

1. *Gain Competitive Advantage*: Use AI-powered analysis of unstructured data to gain insights that drive business innovation and competitiveness.
2. *Improve Operational Efficiency*: Analyze unstructured data to identify areas for process improvement and optimize business operations.
3. *Enhance Customer Engagement*: Use AI-powered analysis of unstructured data to personalize customer experiences and improve engagement.

5) Importance of Research:

The present research work would prove useful and relevant as it would throw light upon This research is important for several reasons:

1. International importance: The proposed system can be used in various countries to improve road safety and reduce the risk of accidents and fatalities.

2. National importance: The proposed system can be used in various industries, such as transportation, logistics, and emergency services.

3. Regional importance: The proposed system can be used in various regions to improve road safety and reduce the risk of accidents and fatalities. The discussed issues will be equally valuable to the scholars, readers and researchers.

6) Scope of Research: The scope of this research includes:

The analysis will focus on the application of AI techniques to unstructured data sources, including:

1. *Text data*: Social media posts, customer reviews, emails, chat logs, and documents.

2. *Image data*: Images, videos, and other visual content from sources like social media, surveillance cameras, and medical imaging.

3. *Audio data*: Audio recordings, voice messages, and phone calls.

AI Techniques

The analysis will explore the following AI techniques:

1. *Natural Language Processing (NLP)*: Text preprocessing, sentiment analysis, entity recognition, and topic modeling.

2. *Machine Learning (ML)*: Supervised, unsupervised, and reinforcement learning techniques for classification, regression, and clustering.

3. *Deep Learning (DL)*: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transfer Learning.

Applications

The analysis will investigate the applications of AI-powered unstructured data analysis in various domains, including:

1. *Customer Service*: Sentiment analysis, intent detection, and personalized recommendations.

2. *Marketing*: Social media monitoring, influencer identification, and targeted advertising.

3. *Healthcare*: Medical imaging analysis, disease diagnosis, and patient outcome prediction.

4. *Finance*: Risk assessment, fraud detection, and portfolio optimization.

7) Hypothesis:

Primary Hypotheses

1. The application of AI techniques to unstructured data will result in improved accuracy and efficiency in data analysis compared to traditional methods.

2. The use of natural language processing (NLP) techniques will enable the extraction of meaningful insights from unstructured text data.

3. The integration of machine learning (ML) and deep learning (DL) algorithms will enhance the predictive power of unstructured data analysis.

Secondary Hypotheses

1. The analysis of unstructured data using AI techniques will reveal hidden patterns and relationships that are not apparent through traditional analysis methods.

2. The use of AI-powered unstructured data analysis will lead to improved decision-making and strategic planning in organizations.

3. The application of AI techniques to unstructured data will enable the identification of new business opportunities and revenue streams.

Domain-Specific Hypotheses

1. In the healthcare domain, the analysis of unstructured clinical notes using AI techniques will improve disease diagnosis and patient outcome prediction.

2. In the marketing domain, the use of AI-powered unstructured data analysis will enhance customer segmentation and targeted advertising.

3. In the finance domain, the application of AI techniques to unstructured financial data will improve risk assessment and portfolio optimization.

Methodology-Specific Hypotheses

1. The use of supervised learning algorithms will result in better performance than unsupervised learning algorithms for unstructured data analysis.

2. The integration of transfer learning and domain adaptation techniques will improve the accuracy of AI models for unstructured data analysis.

3. The use of ensemble methods will enhance the robustness and reliability of AI-powered unstructured data analysis.

These hypotheses can serve as a starting point for your analysis, and you can refine or expand them based on your specific research questions and objectives.

8) Research Methodology: The research methodology includes:

Research methodology for the analysis of unstructured data using artificial intelligence (AI) techniques:

Research Design

1. *Mixed-Methods Approach*: Combine quantitative and qualitative methods to analyze unstructured data.

2. *Exploratory Study*: Explore the application of AI techniques to unstructured data analysis.

3. *Case Study Research*: Conduct in-depth analysis of real-world cases or datasets.

Data Collection

1. *Secondary Data*: Collect existing unstructured data from various sources (e.g., social media, text documents, images).

2. *Primary Data*: Collect new unstructured data through surveys, interviews, or observations.

3. *Data Preprocessing*: Clean, preprocess, and transform data into a suitable format for AI analysis.

AI Techniques

1. *Natural Language Processing (NLP)*: Use NLP techniques (e.g., text preprocessing, sentiment analysis, entity recognition) for text data analysis.

2. *Machine Learning (ML)*: Apply ML algorithms (e.g., supervised, unsupervised, reinforcement learning) for pattern recognition, classification, and prediction.

3. *Deep Learning (DL)*: Utilize DL techniques (e.g., convolutional neural networks, recurrent neural networks) for image, audio, and text analysis.

Data Analysis

1. *Descriptive Statistics*: Calculate summary statistics (e.g., mean, median, standard deviation) for quantitative data.

2. *Inferential Statistics*: Apply statistical tests (e.g., hypothesis testing, regression analysis) to identify patterns and relationships.

3. *AI Model Evaluation*: Assess the performance of AI models using metrics (e.g., accuracy, precision, recall, F1-score).

Validation and Reliability

1. *Cross-Validation*: Use techniques (e.g., k-fold cross-validation) to evaluate AI model performance on unseen data.

2. *Model Interpretability*: Analyze AI model results to understand the relationships between variables and the decision-making process.

3. *Human Evaluation*: Have human evaluators assess the accuracy and relevance of AI-generated insights.

Ethical Considerations

1. *Data Privacy*: Ensure the confidentiality and anonymity of data sources.

2. *Bias and Fairness*: Address potential biases in AI models and ensure fairness in decision-making processes.

3. *Transparency and Accountability*: Provide transparent explanations of AI model results and ensure accountability for AI-driven decisions.

Timeline

1. *Literature Review*: 2 weeks

2. *Data Collection and Preprocessing*: 4 weeks

3. *AI Model Development and Evaluation*: 8 weeks

4. *Results Interpretation and Validation*: 4 weeks

5. *Writing and Revisions*: 8 weeks

This research methodology provides a general outline for analyzing unstructured data using AI techniques. You can refine and adapt it to suit your specific research questions and objectives

9) Proposed Research Plan

i) Plan of Work and target to be achieved in the first year

During the first year an effort will be made to submit the two progress report at the end of the year. This is the beginning of research so go through the various other articles, thesis and books. Try to collect information regarding the topic. Try to understand the things which are going to use while doing research.

ii) Plan of Work and target to be achieved in the second year

During the second year by making use of algorithms along with database positive and negative association rules will be found out and focus wil be given to creating a wireless alarm system to detect nearby moving vehicles is a fascinating project that combines multiple fields of technology, including sensors, wireless communication, and embedded systems.

iii) Plan of Work and target to be achieved in the third year

During the third year thesis writing will be started and publication of two research papers in the reputed journals will be undertaken. Also the research work will be presented in national and international seminar and conference.

10) Expected Outcomes:

The expected outcomes of this research are:

1. Come to know unlock hidden patterns, relationships, and meaning from unstructured data using AI-powered tools.
2. Provide actionable recommendations to stakeholders by analyzing unstructured data and identifying trends, opportunities, and risks.
3. Enhance Customer Experience and analyze customer feedback, sentiment, and behavior from unstructured data sources to improve customer satisfaction and loyalty.
4. Analyzed unstructured data to identify inefficiencies, bottlenecks, and areas for process optimization.
5. Used machine learning and deep learning techniques to develop predictive models that forecast future trends and outcomes.
6. Identified potential risks, anomalies, and threats from unstructured data sources to improve risk management and compliance.

11) Limitations:

The limitations of this research are:

1. It is limited to Artificial Intelligence techniques.
2. Use techniques (e.g., k-fold cross-validation) to evaluate AI model performance on unseen data.
3. AI model results to understand the relationships between variables and the decision-making process
4. Have human evaluators assess the accuracy and relevance of AI-generated insights.

References:

- "Structured vs Unstructured Data | Semi Structured Data." https://k21academy.com/microsoft-azure/dp900/structured-data-vs-unstructured-data-vs-semi-structured-data/ (accessed Dec. 05, 2022).
- What is Unstructured Data | How it is stored | Sources | Real Life Examples | Amit Thinks | 2022, (Apr. 10, 2022).
- Cohen, W. W., et al. (2015). Unstructured data: A survey. ACM Computing Surveys, 47(4), 1-36.
- Kaplan, A. M., et al. (2010). Users of the world, unite! The challenges and opportunities of social media. Business Horizons, 53(1), 59-68.
- Khan, S., et al. (2019). Data quality issues in big data. IEEE Transactions on Industrial Informatics, 15(4), 1784-1793.
- Kotsiantis, S. B., et al. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.
- LeCun, Y., et al. (2015). Deep learning. Nature, 521(7553), 436-444.
- Lipton, Z. C., et al. (2018). The mythos of model interpretability. ACM Queue, 16(3), 31-54.
- Liu, B., et al. (2019). Sentiment analysis: A survey. IEEE Transactions on Affective Computing, 10(2), 147-163.
- Nelson, D., et al. (2017). Document analysis: A survey. ACM Computing Surveys, 50(2), 1-36.
- Rajaraman, A., et al. (2018). Scalable data science. ACM Transactions on Management Information Systems, 9(2), 1-35.

Researcher

Miss Pooja Sanjay Malhare

Research Supervisor

Dr. Purushottam Anandrao Kadam